Preservation of Newspapers

Preliminary Findings of the *Chronicles in Preservation* Project

Katherine Skinner, Executive Director, Educopia Institute
Martin Halbert, Dean of Libraries, University of North Texas
Matt Schultz, Program Manager, MetaArchive Cooperative
Presentation Overview

- Chronicles Project background
- State of the Field report
- Phase 1 Findings and Progress
The Calf-Path by Sam Walter Foss

One day, through the primeval wood,
A calf walked home, as good calves should;
But made a trail all bent askew,
A crooked trail, as all calves do.
Since then three hundred years have fled,
And, I infer, the calf is dead.
But still he left behind his trail,
And thereby hangs my moral tale.
The trail was taken up next day
By a lone dog that passed that way;
And then a wise bellwether sheep
Pursued the trail o’er vale and steep,
And drew the flock behind him, too,
As good bellwethers always do.
And from that day, o’er hill and glade,
Through those old woods a path was made,
And many men wound in and out,
And dodged and turned and bent about,
And uttered words of righteous wrath
Because ’twas such a crooked path;
But still they followed — do not laugh —
The first migrations of that calf.

This crooked path became a road,
Where many a poor horse with his load,
Toiled on beneath the burning sun,
And traveled some three miles in one.
And thus a century and a half,
They trod the footsteps of that calf.
The years passed on in swiftness fleet,
The road became a village street;
And this, before men were aware,
A city’s crowded thoroughfare;
And soon the central street was this,
Of a renowned metropolis;
And men two centuries and a half,
Trod the footsteps of that calf.
Each day, a hundred thousand men were led
By one calf near three centuries dead.
They follow still his crooked way,
And lose one hundred years a day,
For thus such reverence is lent
To well-established precedent.
Chronicles, an introduction

Educopia Institute-led partnership, comprised of the following:

**Preservation groups**
- MetaArchive (LOCKSS)
- Chronopolis (iRODS)
- University of North Texas (CODA)

**Content Curators**
- Penn State
- Virginia Tech
- University of Utah
- Georgia Tech
- Boston College
- Clemson University
- University of Kentucky
- University of North Texas
Project aims

to study, document, and model the use of data preparation practices and distributed digital preservation frameworks to collaboratively preserve digitized and born-digital newspaper collections.
Three DDP frameworks

- **MetaArchive**
  - Founded 2004, 50+ members in 4 countries
  - Multi-node, wide distribution of content

- **Chronopolis**
  - 3-node system (UCSD, SDSC, NCAR, UMIACS)

- **CODA**
  - Developing multi-node framework based on a micro-services approach
Two types of content

Born Digital

Digitized
Research Questions

- How can curators effectively and efficiently prepare their existing digitized and born-digital newspaper collections for preservation?

- How can curators ingest preservation-ready newspaper content into existing DDP solutions?

- What are the strengths and challenges of three leading DDP solutions when used to preserve digital newspaper content?
Project Deliverables

- Guidelines to Digital Preservation Readiness
- Chronicles Repository Interoperability Tools
- Comparative Analysis of DDP Frameworks
State of the Field

- Early findings based on the following surveys:
  - 2009 Digital Preservation Needs Survey (NHPRC)
  - 2011 Digital Preservation SPEC Kit 325 (ARL)
  - 2012 Chronicles Survey (>60 newspaper archives)
  - 2011-12 Chronicles In-Depth Interviews (academic libraries)
State of the field: 2011-12

Chronicles Project Survey

Type
- NDNP: 71; non-NDNP: 495; born digital: 39

Image/derivative formats
- TIFF, JP2, PDF, HTML, TXT, XML

Metadata formats
- METS/ALTO, MIX, MODS, PREMIS

OCR formats
- METS, ALTO, PDF, Abbyy, XML, PRIME OCR.pro
State of the field: 2011-12

- Chronicles Project Survey (cont)
  - Object identifier schemes
    - Fedora PID, Handles, Veridian and CONTENTdm custom URLs, ARKs, NOID and CDL Identity
    - All but two are internal to the repository system
  - Validation
    - \( \frac{1}{2} \) use JHOVE at least for some content
  - Versioning
    - Only one institution
Survey Findings, general

- Back-up vs. preservation
- Adoption of existing standards is low
  - e.g., OCR, metadata
- Lack of standards
  - e.g., file structures, naming conventions, and object identifier schemes
- Diverse array of expectations for access & recovery
  - very institution-specific
- Versioning processes will be necessary
  - e.g., for growing, changing, and/or remediated projects
Guidelines to Digital Newspaper Preservation Readiness

- Background and Stakeholders
- Essential/Optimal Guidelines
  - Acquisition
  - Selection
  - Persistent Identifiers
  - Metadata
  - Data Structures & Repository Frameworks
  - Documentation of Local Conventions
  - IP Considerations
  - Costs
  - Vendor Recommendations
  - “Data Wrangling”
Guidelines: Challenges

• Audience (giving enough information without overwhelming)
• Distinguishing between born-digital and digitized content
• Including case studies of “real world” operations
  • Today’s workflows
  • Remediation workflows
  • Policies/planning

Guidelines for Digital Newspaper Preservation Readiness

Version 0.1, September 10, 2012
Draft Version – For Comment

Katherine Skinner, Educopia Institute
Matt Schultz, MetaArchive Cooperative
Cathy Hartman, University of North Texas
Martin Halbert, University of North Texas
Hannah Tarver, University of North Texas
Mark Phillips, University of North Texas
Gail McMillan, Virginia Tech
David Minor, Chronopolis, SDSC
Mary Molinaro, University of Kentucky
Sue Kellerman, Pennsylvania State University
Robert Horton, Institute for Museum and Library Services
Liz Bishoff, The Bishoff Group
Frederick Zarnold, Global Connexions
Interoperability Tools

- BagIt (documenting optional fields for bag-info.txt file)
- Digital Object Manifest Manager
- METS Newspaper Profile Schemas
- DAITSSS Description Service
- PREMIS Event Tracker API
Questions we have for you

- What type of “guidelines” would be most useful in your environment?
- Do your practices resemble what we’ve discussed.
Contact

Katherine Skinner
katherine.skinner@metaarchive.org

Martin Halbert
martin.halbert@unt.edu

Matt Schultz
matt.schultz@metaarchive.org